Journal of Informetrics 11 (2017) 625–640

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

The Herrero-Villar approach to citation impact

Pedro Albarrán[a], Carmen Herrero[b], Javier Ruiz-Castillo[c,*], Antonio Villar[d]

[a] Departamento de Fundamentos del Análisis Económico, Universidad de Alicante, Spain
[b] Departamento de Fundamentos del Análisis Económico, Universidad de Alicante & IVIE, Spain
[c] Departamento de Economía, Universidad Carlos III, Spain
[d] Departamento de Economía, Universidad Pablo de Olavide & IVIE, Spain

ARTICLE INFO

Article history:
Received 19 December 2016
Received in revised form 26 April 2017
Accepted 26 April 2017

Keywords:
Citation impact
Additive rules
Tournaments
Relative research performance
Worth

ABSTRACT

This paper focuses on the evaluation of research institutions in terms of size-independent indicators. There are well-known procedures in this context, such as what we call additive rules, which provide an evaluation of the impact of any research unit in a scientific field based upon a partition of the field citations into ordered categories, along with some external weighting system to weigh those categories. We introduce here a new ranking procedure that is not an additive rule – the HV procedure, after Herrero & Villar (2013) – and compare it those conventional evaluation rules within a common setting. Given a set of ordered categories, the HV procedure measures the performance of the different research units in terms of the relative probability of getting more citations. The HV method also provides a complete, transitive and cardinal evaluation, without recurring to any external weighting scheme. Using a large dataset of publications in 22 scientific fields assigned to 40 countries, we compare the performance of several additive rules – the Relative Citation Rate, four percentile-based ranking procedures, and two average-based high-impact indicators – and the corresponding HV procedures under the same set of ordered categories. Comparisons take into account re-rankings, and differences in the outcome variability, measured by the coefficient of variation, the range, and the ratio between the maximum and minimum index values.

1. Introduction

In a globalized and highly interconnected world, comparative exercises have become more and more frequent in many aspects of life. Research is no exception and there seems to be growing interest in the evaluation of the scientific influence. Citation analysis has become one of the key tools for evaluating the scientific performance of research units (individual authors, research groups, departments, universities, countries, etc.). Citation impact indicators differ depending on the evaluation approach, the motivation, and the way of transforming citations into specific evaluation formulae. In this paper, we contribute to the literature of citation analysis that focuses on the *ranking of research units* by *size-independent measures of citation impact* when size is measured by the number of publications, i.e. measures that take the relative citation frequencies as the basis for the evaluation (see, among many others, Bornmann, De Moya Anegón, & Leydesdorff, 2012, Fairclough & Thelwall, 2015, and Glänzel, Thijs, & Debackere, 2014). We propose a new procedure that evaluates the citation impact of a set of research units according to the criterion pioneered in Herrero & Villar (2013) (*HV* in the sequel).

* Corresponding author.
E-mail address: jrc@eco.uc3m.es (J. Ruiz-Castillo).

http://dx.doi.org/10.1016/j.joi.2017.04.008
1751-1577/

This new evaluation protocol can be thought of as a two-step procedure in which we first define a partition of the range of citations into a series of categories that gather publications of similar merit, and then evaluate the research units' relative citations distributions embedded in these categories. The key informational item to compare research units is, therefore, the shares of the publications into the different categories. The comparison of these distributions is made in terms of the following principle: each research unit will be compared with all others in terms of the probability of getting a greater citation impact. We shall see that this procedure can also be formulated in terms of a series of tournaments in which each research unit is confronted with all others repeatedly.[1]

There are well-established evaluation procedures that also rely on the assessment of the research units' relative citations distributions by categories using different principles. We shall consider here three types of these indicators that will be used as reference for comparison with the *HV* evaluations. The first type is the *Relative Citation Rate, RCR* (Schubert & Braun, 1986, and Vinkler, 1986). The second type, promoted since 2010 by a group of highly qualified professional leaders in scientometrics, corresponds to what Bornmann & Mutz (2011) call the *percentile rank approach*.[2] The third type consists of the FGT family of high-impact indicators, introduced in Albarrán, Ortuño, and Ruiz-Castillo (2011a), which are real valued functions defined over the subset of publications with citations above a *critical citation line* (CCL), and whose properties are inherited from a class of economic poverty indicators introduced by Foster, Greeer and Thorbecke (1984).[3]

All these evaluation procedures have in common an additive structure, as shown in Section 2. That is, the evaluation of the different research units is given by a weighted sum of the relative citations distribution by categories, where the weights measure the importance of each category. They can be described as implementing the following protocol. First, publications are distributed into a set of categories that gather those publications regarded as being of similar merit. Second, each of those categories is given a weight that determines the rate of substitution between the corresponding categories. And third, the evaluation of each research unit is obtained as a weighted sum of its relative frequencies aggregated by categories.

The additive structure of these evaluation procedures is very appealing because it provides a relatively simple construct that is easy to interpret and rather immediate to compute. The main shortcoming of these indicators is that the evaluation turns out to depend critically on the choice of the weights with which we ponder the publications. Quite often there is no good reason to choose a particular weighting system, which makes the evaluation exercise somehow arbitrary because both the evaluation of the individual units and their ranking depend on those weights. The new evaluation approach presented in this paper avoids this inconvenience because no weighting of categories is involved and still provides a complete, transitive and cardinal evaluation.

Any comparison between alternative ranking procedures should involve not only their rationale and their properties but also the empirical differences they give rise to in applications. Following this idea, we consider here an empirical analysis based on a dataset indexed by Thomson Scientific, and consisting of 4.4 million articles published in 1998–2003, and the citations they received during a five-year citation window for each year in that period. Articles are classified into the 20 natural sciences and the two social sciences distinguished by this firm. Using these data, we compare the *HV* ranking procedure and the ranks provided by a group of additive procedures in four scientific fields.[4] We compare ranking procedures both from an ordinal point of view (changes in the ranking) and from a cardinal point of view (differences in the spread of the evaluations, as measured by the coefficient of variation, CV, the range, and the ratio between the maximum and minimum). We study a partition of the world into 39 countries and a residual geographical area.

The remainder of this paper is organized into four Sections and an Appendix. Section 2 presents a selection of additive ranking procedures. Section 3 describes the alternative approach for the evaluation of research units in a single field by adapting the ideas in Herrero & Villar (2013) to this context. The empirical Section 4 develops a comparison between this new ranking procedure and the selected additive procedures in the following fields: Clinical Medicine, Physics, Engineering, and Economics & Business. These fields have been selected endeavoring to ensure diversity and relevance, while keeping the set of empirical comparisons within reasonable limits. Section 5 contains discussion. The Appendix A includes some examples and descriptive statistics. Also, to facilitate reading of the text, some statistical results are relegated to a Supplementary Material section.

[1] The recourse to tournaments as an evaluation procedure has been applied in related contexts, such as the Google Page Rank algorithm to rank web pages (Altman & Tennenholtz, 2005 and Page et al., 1998), as well as the invariant method (Palacios-Huerta and Volij, 2004), the Eigenfactor (Bergstrom, 2007, and West et al., 2010), and the recent paper by Kóczy & Nichifor (2013) that have been used to rank scientific journals. The closest contribution to ours is Carayol & Lahatte (2014), which uses the idea of tournaments for ranking research units when citation impact and quantity both matter.

[2] See also the Integrated Impact, or the *I3* indicator in Leydesdorff et al. (2011), Leydesdorff and Bornmann (2011), Leydesdorff (2012), Wagner & Leydesdorff (2012), and Rousseau (2012). In their search for standards for applying bibliometric methods in the evaluation of research institutes or individuals, Bornmann & Williams (2013), as well as Bornmann, Marx, and co-authors point to the percentile rank approach as the obvious choice (Bornmann & Marx, 2013, 2014, and Bornmann, Mutz, Nehaus, & Daniel, 2008, Bornmann et al., 2014).

[3] For empirical applications of members of the FGT family, see Albarrán, Ortuño, and Ruiz-Castillo (2011b), Albarrán, Ortuño, and Ruiz-Castillo (2011c), Herranz & Ruiz-Castillo (2012, 2013), and Perianes-Rodriguez & Ruiz-Castillo (2016).

[4] The study has been made for all 22 fields, obtaining similar results. We here report the results for four fields for the sake of parsimony. All remaining results can be obtained from the authors upon request.